BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

REMARKS

Claims 1, 8-16, 60, 77 and 78 have been amended. Claims 3-7, 19-59, 61-63, 66-74 and 76 have been canceled without prejudice or disclaimer. Subsequent to the entry of the present amendment, claims 1, 2, 8-16, 17, 18, 60, 64, 65, 75, 77, 78, 79 and 80 are pending and at issue. These amendments and additions add no new matter as the claim language is fully supported by the specification and original claims.

Applicants note that some of the claims as amended recite a single chain antibody having the amino acid sequence set forth in SEQ ID NO:2. The sequence listing filed April 27, 2001 clearly indicates that the amino acid sequence of SEQ ID NO:2 is encoded by the nucleic acid sequence of SEQ ID NO:1. Reconsideration of the application in light of the foregoing amendments and the following discussion is respectfully requested.

Claim Objections

Claims 9-16 have been objected to as depending from a canceled claim. The claims have been amended to depend from a pending claim. Thus, this rejection is now moot.

The Rejection under 35 U.S.C. § 112, First Paragraph, Enablement

Claims 1-2, 5-6, 8-18, 60, 63-65, 75-76 and 79-80 stand rejected under 35 U.S.C. § 112, first paragraph as allegedly not enabled by the specification as filed. This rejection is moot with regard to canceled claims 5-6, 63 and 76. Applicant respectfully traverses the rejection as it may apply to the amended claims.

The Office Action acknowledges that the present specification enables a variety of embodiments, as listed in points 1 to 15, starting on page 2, item 4, of the Office Action, including methods that include a single chain antibody encoded by SEQ ID NO:1 that binds specifically to phOx and a linker coupling a probe to a ligand such as phOx. However, the Office Action alleges that the specification does not enable a method that includes *any* single chain antibody, including a single chain antibody that has at least 30% sequence identity to SEQ ID NO:1, that binds to *any* ligand (emphasis in original). Claims 1 and 60 have been

amended to recite a single chain antibody that specifically binds to phOx. In addition, claims 1 and 60 have been amended to recite a ligand comprising phOx. The present disclosure provides working examples of methods for localizing a probe that includes an antibody that binds phOx. The specific single chain antibody/ligand combination functions in the method to localize the probe to the vicinity of the ligand, allowing visualization of the ligand. Accordingly, the claimed methods should not be limited to solely to the use of a single chain antibody having the amino acid sequence set forth in SEQ ID NO:2. It would be a matter of routine experimentation, not undue experimentation, for one skilled in the art to identify a single chain antibody that specifically binds to phOx and possesses an amino acid sequence different from SEQ ID NO:2.

Regarding undue experimentation, as stated by the Patent Office, the Federal Circuit in In re Wands directed that the focus of the enablement inquiry should be whether the experimentation needed to practice the invention is or is not "undue" experimentation. The court set forth specific factors to be considered.

One of these factors is "the quantity of experimentation necessary." Guidance as to how much experimentation may be needed and still not be "undue" is set forth by the Federal Circuit in, e.g., Hybritech, Inc. v. Monoclonal Antibodies, Inc. In that case, an applicant had claims that were generic to all IgM antibodies directed to a specific antigen. However, only a single antibody producing cell line had been deposited. The PTO had rejected claims that were generic to all antibodies directed to the antigen as lacking an enabling disclosure.

The Federal Circuit reversed, noting that the evidence indicated that those skilled in the monoclonal antibody art could, using the state of the art and applicants' written disclosure, produce and screen new hybridomas secreting other monoclonal antibodies falling within the genus without undue experimentation. The court held that applicants' claims need not be limited to the specific, single antibody secreted by the deposited hybridoma cell line (significantly, the genus of antibodies was allowed even though only one antibody species was disclosed). The court was acknowledging that, because practitioners in that art are prepared to screen large numbers of negatives in order to find a sample that has the desired properties, the

screening that would be necessary to make additional antibody species was not "undue experimentation."

Analogously, practitioners of molecular biology for the instant invention also recognize that many rounds of screening may be necessary to identify and isolate single chain antibodies that specifically bind to phOx. However, the procedures for isolating such antibodies are widely accepted, routine protocols, not requiring "undue experimentation" to be practiced. Accordingly, one skilled in the art has sufficient guidance by the specification to practice the claimed methods without undue experimentation.

A skilled artisan can practice the present invention using standard research techniques for isolating single chain antibodies with a particular binding specificity. The level of skill and knowledge in the art is exemplified by patents, that were filed before the filing date of the instant application, claiming an expression vector comprising a DNA sequence encoding a single chain antibody. For example, in U.S. Patent No. 6,017,754, claim 1 reads in part:

> 1. A eukaryotic expression vector ... comprising:
> a first DNA sequence encoding an anti-hapten single-chain antibody, which antibody binds to a specific hapten, wherein said hapten is 4-ethoxymethylene-2-phenyl-2-oxazolin-5-one (i.e., phOx).[1]

Applicants respectfully submit that the written disclosure of the instant application is supplemented by the knowledge held by one of ordinary skill in the art. The skilled artisan is one who is knowledgeable about basic laboratory/research protocols. It is well settled law that an Applicant need not include disclosure that was well known in the art. Furthermore, in addition to the knowledge held by the skilled artisan, Applicants provide exemplary basic

---

[1] U.S. Patent No. 6,017,754, column 33, lines 40-44.

regarding a single chain antibody that can be used in the methods of the invention (i.e., one that specifically binds to phOx). Applicants respectfully submit that the application, at the time of filing, taught one of skill in the art how to practice the claimed method.

In addition, Applicants have amended claim 77 to recite a single chain antibody that comprises:

a) the amino acid sequence set forth in SEQ ID NO:2;
b) the amino acid sequence set forth in SEQ ID NO:2 with up to 30 conservative amino acid substitutions;
c) an amino acid sequence at least 95% identical to SEQ ID NO:2;
d) an amino acid sequence encoded by the nucleic acid sequence set forth in SEQ ID NO:1; or
e) an amino acid sequence encoded by a nucleic acid sequence at least 95% identical to SEQ ID NO:1.

Applicants have also added new claim 81 which recites a binding partner encoded by:

a) a nucleic acid sequence comprising SEQ ID NO:1;
b) a nucleic acid sequence at least 95% identical to SEQ ID NO:1;
c) a nucleic acid sequence encoding a polypeptide consisting of the amino acid sequence set forth in SEQ ID NO:2 with up to 30 conservative amino acid substitutions; or
d) a nucleic acid sequence encoding a polypeptide consisting of an amino acid sequence at least 95% identical to SEQ ID NO:2.

Support for the new claims can be found beginning at page 11, line 11, bridging to page 13, line 23. While the Office Action has not rejected claim 77 as amended, or new claim 81, the Office Action does assert that the specification provides insufficient guidance as to which amino acids, and the corresponding nucleotides within the full length sequence of SEQ ID NO:1, can be modified such that the resulting single chain antibody maintains the same binding specificity as the antibody encoded by SEQ ID NO:1. To support this assertion, the Examiner cites three references: Skolnick et al., Ngo et al., and Abaza et al. The Examiner appears to take the position that these references demonstrate that even a single amino acid substitution or 'conservative' amino acid substitution in a protein will often dramatically affect

the biological activity and characteristics of a protein, and concludes that undue experimentation would be required to enable the full scope of the claims. Applicants respectfully disagree.

Applicants agree that it is possible, at least in some cases, to abolish activity of a given protein by mutating a critical residue, as disclosed by the cited references. However, applicants disagree that this fact means that one of ordinary skill cannot make functional analogs of SEQ ID NO:2 without undue experimentation. In support of this, Applicants provide EXHIBIT A (Bowie et al., Science 247:1305). Bowie et al. teaches, at page 1306, col.2, lines 12-13, that "proteins are surprisingly tolerant of amino acid substitutions." Bowie et al. cites as evidence a study carried out on the lac repressor. Of approximately 1500 single amino acid substitutions at 142 positions in this protein, about one-half of the substitutions were found to be "phenotypically silent": that is, had no noticeable effect on the activity of the protein (page 1306, col. 2, lines 14-17). Presumably the other half of the substitutions exhibited effects ranging from slight to complete abolishment of repressor activity. Thus, one can expect, based on Bowie et al.'s teachings, to find over half (and possibly well over half) of random substitutions in any given protein to result in mutated proteins with full or nearly full activity. These are far better odds than those at issue in In re Wands, 858 F.2d 731 (Fed. Cir. 1988), in which the court said that screening many hybridomas to find the few that fell within the claims was not undue experimentation. The question is not whether it is possible to abolish activity with a modification such as a point mutation, but rather whether one of ordinary skill can produce, without undue experimentation, modified single chain antibodies in which the activity of specifically binding to phOx is not abolished. Based on Bowie et al.'s teachings, one would predict that even random substitution of residues in SEQ ID NO:2 will predictably result in a majority of the modified antibodies having full or partial phOx binding activity.

In view of the amendments to the claims, and in light of the above discussion, Applicants request withdrawal of the rejection of claims 1-2, 8-18, 60, 64-65, 75 and 79-80 under 35 U.S.C. § 112, first paragraph.

The Rejection under 35 U.S.C. § 112, First Paragraph, Written Description

Claims 1-3, 5-6, 8-18, 60, 63-65, 75-76 and 79-80 stand rejected under 35 U.S.C. § 112, first paragraph as allegedly not adequately described by the specification. This rejection is moot with regard to canceled claims 3, 5-6, 63 and 76. Applicant respectfully traverses the rejection as it may apply to the amended claims.

The Office Action alleges that because the specification discloses only one single chain antibody encoded by the nucleic acid sequence set forth in SEQ ID NO:1, it does not sufficiently describe methods that include any single chain antibody that binds to any ligand (citing to *University of California v. Eli Lilly and Co.*, 43 USPQ 2d, 1398; and *University of Rochester v. G.D. Searle & Co.*, 69 USPQ2d 1886). As previously noted, claims 1 and 60 have been amended to recite a single chain antibody that specifically binds to phOx. In addition, claims 1 and 60 have been amended to recite a ligand comprising phOx. In view of the support for the amended claims provided in the specification, Applicants maintain that the claimed methods should not be limited to solely to the use of a single chain antibody having the amino acid sequence set forth in SEQ ID NO:2.

The present disclosure provides working examples of methods for localizing a probe that include a single chain antibody that binds phOx. While the claims as amended encompass the use of a genus of single chain antibodies that bind to phOx, Applicants note that the law does not require that the specification describe every species within the genus. As described above, Applicants have provided at least one amino acid sequence of a member of the genus of single chain antibodies used in the claimed methods. Applicants have further provided relevant identifying characteristics of the genus encompassed by the claims (i.e., a single chain antibody that binds to phOx). Accordingly, Applicants have clearly demonstrated that they were in "possession of the necessary common attributes of features of the elements possessed by members of the genus" (66 Fed. Reg. 1099, at 1106) as of the filing date of the application.

In summary, Applicant respectfully requests withdrawal of the rejection of the claims under 35 U.S.C. § 112, first paragraph as allegedly not adequately described by the specification as filed.

The Rejection under 35 U.S.C. § 103

Claims 60 and 63 stand rejected under 35 U.S.C. § 103(a) as allegedly unpatentable over U.S. Pat. No. 6,017,754 (referred to herein as "the '754 patent") in view of Haugland et al. (Handbook of Fluorescent Probes and Research Chemicals 6th edition, pages 13-15, 18-19 (1996)) and WO 93/11120. This rejection is moot with regard to canceled claim 63. Applicants traverse this rejection as it may apply to the amended claims.

Applicants note that claim 60 has been amended to recite a single chain antibody that specifically binds to phOx. In addition, claim 60 has been amended to recite a ligand comprising phOx. Finally, claim 60 has been amended to recite the step of "detecting the probe/ligand conjugate within the cell, thereby localizing the probe within the cell." The combination of the '754 patent and the cited secondary references, does not result in a method for localizing a probe within a cell, that includes a membrane permeant conjugate for detecting a single chain antibody or a specific binding pair member expressed from a recombinant nucleic acid, as recited in claim 60.

In view of the fact that none of the cited references, alone or in combination, teach or suggest a method for localizing a probe inside a cell, Applicants request withdrawal of this rejection.

Claims 1-2, 5-6, 8, 11-14, 16-17, 64-65 and 75-76 stand rejected under 35 U.S.C. 103(a) as allegedly unpatentable over the '754 patent (US Pat No. 6,017,754) in view of Schouten et al, Haugland et al. (Handbook of Fluorescent Probes and Research Chemicals 6th edition, 1996, pages 13-15, 18-19) and WO 93/11120. This rejection is moot with regard to canceled claims 5-6 and 76. Applicant respectfully traverses the rejection as it may apply to the amended claims.

The Office Action alleges that the '754 patent teaches a method of identifying and selecting a cell to study genes of interest at a cellular level by transfecting the cell with a plasmid that encodes a single chain antibody (sFv) directed against phOx. The Office Action further asserts that the '754 patent teaches that the hapten (phOx) as the ligand can be

conjugated to a fluorescent (FITC) spectroscopic probe or other label via a linker moiety (phOx-BSA-FITC) to allow for identification and selection of the transfected cell by detecting fluorescence emission (citing column 7, line 8-13 of the '754 patent). Schouten et al has been added to the list of references cited in previous Office Actions. Schouten allegedly teaches a method of targeting a single chain antibody to various subcellular locations by fusing various targeting signals to the antibody. Haugland et al. allegedly teach spectroscopic probes that are membrane permeant, including BODIPY FL. WO 93/11120 allegedly teaches a flexible aliphatic linker that is membrane permeant. Based on these assertions, the Office Action concludes that it would have been obvious to one of ordinary skill in the art at the time the invention was made to combine 1) the method of using a phOx-binding single chain antibody to identify and isolate cells, as taught in the '754 patent; 2) the single chain antibody containing a subcellular localization signal, as taught by Schouten; 3) the impermeant linker/probe conjugate taught by Haugland; and 4) the linker taught by Haugland or by WO 93/111020, to arrive at the presently claimed method of localizing a signal inside a cell.

When a rejection depends on a combination of prior art references, there must be some teaching, suggestion, or motivation to combine the references. The Court of Appeals for the Federal Circuit has restated the general principle that hindsight analysis cannot be a basis for an obviousness rejection:

> The [Patent Office] did not, however, explain what specific understanding or technological principle within the knowledge of one of ordinary skill in the art would have suggested the combination [of references cited]. Instead, the [Patent Office] merely invoked the high level of skill in the field of art. If such a rote invocation could suffice to supply a motivation to combine, the more sophisticated scientific fields would rarely, if ever, experience a patentable technical advance. Instead, in complex scientific fields, the [Patent Office] could routinely identify the prior art elements in an application, invoke the lofty level of skill, and rest its case for rejection. To counter this potential weakness in the

> obviousness construct, the suggestion to combine requirement stands as a critical safeguard against hindsight analysis and rote application of the legal test for obviousness.[2]

In the instant Office Action, the Patent Office fails to sufficiently explain what specific understanding or technological principle within the knowledge of one of ordinary skill in the art would have suggested the combination of the four cited references to arrive at Applicants claimed invention. Simply stating that one of ordinary skill in the art would be motivated to make the combination because: 1) single chain antibodies are encoded by small nucleic acid coding sequences; 2) targeting signals are useful for targeting fusion polypeptides to a particular cellular location; 3) cell permeant probes are known to the skilled artisan; and 4) flexible linkers are also known to the skilled artisan, fails to articulate how the references suggest to the skilled artisan that the combination would result in the presently claimed method.

The Examiner seems to be suggesting that the cited references demonstrate that the invention "could" have been made by one skilled in the art. However, this is not the test of obviousness. To be obvious, an invention must be somehow "taught" by the prior art. In this case, none of the references disclose or suggest a method for localizing a probe within a cell, that includes a membrane permeant conjugate for detecting a single chain antibody or a specific binding pair member expressed from a recombinant nucleic acid, as recited in the pending claims. The Examiner has the burden of explaining how the prior art suggests the claimed subject matter and not simply the general aspects of the invention (e.g., expression of a single chain antibody in a cell, the use of targeting signals, etc.). It is unclear how the references, even if properly combined, render the claimed invention obvious. Accordingly, Applicants request withdrawal of this rejection.

---

[2] In re Rouffet, 149 F.3d 1350 (Fed. Cir. 1998).

The Office Action rejects claim 9 (see page 16, part 12 of the Office Action), claim 10 (see page 17, part 13 of the Office Action), claims 15 and 79-80 (see page 18, part 14 of the Office Action), and claim 18 (see page 20, part 15 of the Office Action) under 35 U.S.C. 103(a) as being unpatentable over the various references. Applicants respectfully traverse these rejections.

Claims 9, 10, 15, 18 and 79-80 ultimately depend from independent claim 1. As discussed above, Applicants believe that amended claim 1 is nonobvious. Applicants submit that if an independent claim is nonobvious under 35 U.S.C. §103, then any claim depending therefrom is nonobvious. *In re Fine*, 837 F.2d 1071, (Fed. Cir. 1988); MPEP §2143.03. Accordingly, Applicants request withdrawal of these rejections under 35 U.S.C. 103(a).

In view of the amendments to the claims and the above remarks, reconsideration and favorable action on all claims is respectfully requested. Should any questions remain in view of this communication, the Examiner is encouraged to call the undersigned so that a prompt disposition of this application can be achieved. Please charge any additional fees, or make any credits, to Deposit Account No. 50-1355.

Respectfully submitted,

Date: July 28, 2004

Michael Reed, J.D., Ph.D.
Reg. No. 45,647
Applicant's Representative
Telephone: (858) 638-6754
Facsimile: (858) 677-1465

Gray Cary Ware & Freidenrich LLP
4365 Executive Drive, Suite 1100
San Diego, CA 92121-2133
USPTO Customer Number 28213

EXHIBIT A

Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions

JAMES U. BOWIE,* JOHN F. REIDHAAR-OLSON, WENDELL A. LIM, ROBERT T. SAUER

n amino acid sequence encodes a message that deter-ines the shape and function of a protein. This message is ghly degenerate in that many different sequences can ıde for proteins with essentially the same structure and tivity. Comparison of different sequences with similar essages can reveal key features of the code and improve ıderstanding of how a protein folds and how it per-rms its function.

THE GENOME IS MANIFEST LARGELY IN THE SET OF PRO-teins that it encodes. It is the ability of these proteins to fold into unique three-dimensional structures that allows them to ıction and carry out the instructions of the genome. Thus, nprehending the rules that relate amino acid sequence to struc-e is fundamental to an understanding of biological processes. cause an amino acid sequence contains all of the information :essary to determine the structure of a protein (*1*), it should be ssible to predict structure from sequence, and subsequently to er detailed aspects of function from the structure. However, both oblems are extremely complex, and it seems unlikely that either l be solved in an exact manner in the near future. It may be ssible to obtain approximate solutions by using experimental data simplify the problem. In this article, we describe how an analysis allowed amino acid substitutions in proteins can be used to uce the complexity of sequences and reveal important aspects of ıcture and function.

ethods for Studying Tolerance to quence Variation

There are two main approaches to studying the tolerance of an ino acid sequence to change. The first method relies on the cess of evolution, in which mutations are either accepted or cted by natural selection. This method has been extremely verful for proteins such as the globins or cytochromes, for which uences from many different species are known (*2–7*). The second roach uses genetic methods to introduce amino acid changes at specific positions in a cloned gene and uses selections or screens to identify functional sequences. This approach has been used to great advantage for proteins that can be expressed in bacteria or yeast, where the appropriate genetic manipulations are possible (*3, 8–11*). The end results of both methods are lists of active sequences that can be compared and analyzed to identify sequence features that are essential for folding or function. If a particular property of a side chain, such as charge or size, is important at a given position, only side chains that have the required property will be allowed. Conversely, if the chemical identity of the side chain is unimportant, then many different substitutions will be permitted.

Studies in which these methods were used have revealed that proteins are surprisingly tolerant of amino acid substitutions (*2–4, 11*). For example, in studying the effects of approximately 1500 single amino acid substitutions at 142 positions in *lac* repressor, Miller and co-workers found that about one-half of all substitutions were phenotypically silent (*11*). At some positions, many different, nonconservative substitutions were allowed. Such residue positions play little or no role in structure and function. At other positions, no substitutions or only conservative substitutions were allowed. These residues are the most important for *lac* repressor activity.

What roles do invariant and conserved side chains play in proteins? Residues that are directly involved in protein functions such as binding or catalysis will certainly be among the most conserved. For example, replacing the Asp in the catalytic triad of trypsin with Asn results in a 10^4-fold reduction in activity (*12*). A similar loss of activity occurs in λ repressor when a DNA binding residue is changed from Asn to Asp (*13*). To carry out their function, however, these catalytic residues and binding residues must be precisely oriented in three dimensions. Consequently, mutations in residues that are required for structure formation or stability can also have dramatic effects on activity (*10, 14–16*). Hence, many of the residues that are conserved in sets of related sequences play structural roles.

Substitutions at Surface and Buried Positions

In their initial comparisons of the globin sequences, Perutz and co-workers found that most buried residues require nonpolar side chains, whereas few features of surface side chains are generally conserved (*6*). Similar results have been seen for a number of protein families (*2, 4, 5, 7, 17, 18*). An example of the sequence tolerance at surface versus buried sites can be seen in Fig. 1, which shows the allowed substitutions in λ repressor at residue positions that are near the dimer interface but distant from the DNA binding surface of the protein (*9*). These substitutions were identified by a functional

authors are in the Department of Biology, Massachusetts Institute of Technology, bridge, MA 02139.

ient address: Department of Chemistry and Biochemistry and the Molecular gy Institute, University of California, Los Angeles, Los Angeles, CA 90024.

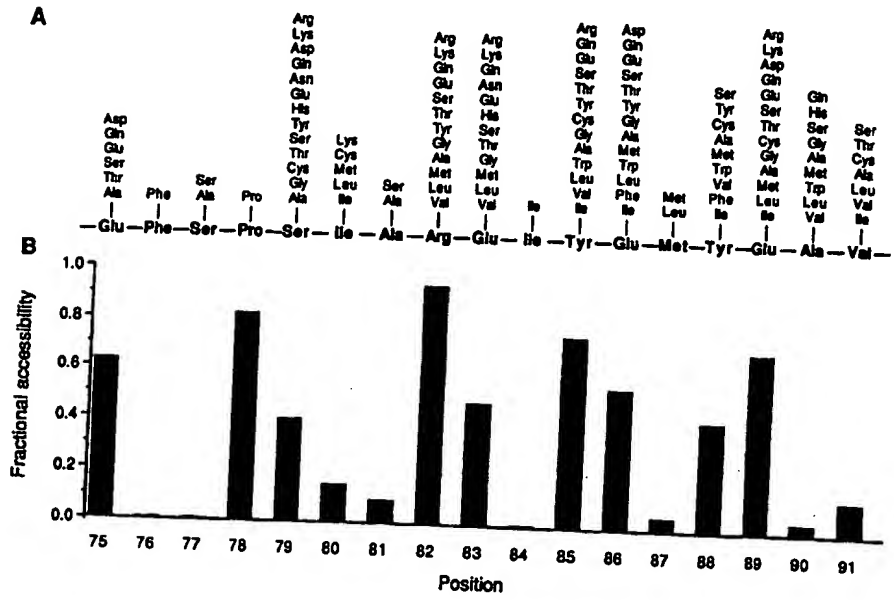

Fig. 1. (**A**) Amino acid substitutions allowed in a short region of λ repressor. The wild-type sequence is shown along the center line. The allowed substitutions shown above each position were identified by randomly mutating one to three codons at a time by using a cassette method and applying a functional selection (*9*). (**B**) The fractional solvent accessibility (*42*) of the wild-type side chain in the protein dimer (*43*) relative to the same atoms in an Ala-X-Ala model tripeptide.

selection after cassette mutagenesis. A histogram of side chain solvent accessibility in the crystal structure of the dimer is also shown in Fig. 1. At six positions, only the wild-type residue or relatively conservative substitutions are allowed. Five of these positions are buried in the protein. In contrast, most of the highly exposed positions tolerate a wide range of chemically different side chains, including hydrophilic and hydrophobic residues. Hence, it seems that most of the structural information in this region of the protein is carried by the residues that are solvent inaccessible.

Constraints on Core Sequences

Because core residue positions appear to be extremely important for protein folding or stability, we must understand the factors that dictate whether a given core sequence will be acceptable. In general, only hydrophobic or neutral residues are tolerated at buried sites in proteins, undoubtedly because of the large favorable contribution of the hydrophobic effect to protein stability (*19*). For example, Fig. 2 shows the results of genetic studies used to investigate the substitutions allowed at residue positions that form the hydrophobic core of the NH_2-terminal domain of λ repressor (*20*). The acceptable core sequences are composed almost exclusively of Ala, Cys, Thr, Val, Ile, Leu, Met, and Phe. The acceptability of many different residues at each core position presumably reflects the fact that the hydrophobic effect, unlike hydrogen bonding, does not depend on specific residue pairings. Although it is possible to imagine a hypothetical core structure that is stabilized exclusively by residues forming hydrogen bonds and salt bridges, such a core would probably be difficult to construct because hydrogen bonds require pairing of donors and acceptors in an exact geometry. Thus the repertoire of possible structures that use a polar core would probably be extremely limited (*21*). Polar and charged residues are occasionally found in the cores of proteins, but only at positions where their hydrogen bonding needs can be satisfied (*22*).

The cores of most proteins are quite closely packed (*23*), but some volume changes are acceptable. In λ repressor, the overall core volume of acceptable sequences can vary by about 10%. Changes at individual sites, however, can be considerably larger. For example, as shown in Fig. 2, both Phe and Ala are allowed at the same core position in the appropriate sequence contexts. Large volume changes at individual buried sites have also been observed in phylogenetic studies, where it has been noted that the size decreases and increases at interacting residues are not necessarily related in a simple complementary fashion (*5, 7, 17*). Rather, local volume changes are accommodated by conformational changes in nearby side chains and by a variety of backbone movements.

The Informational Importance of the Core

With occasional exceptions, the core must remain hydrophobic and maintain a reasonable packing density. However, since the core is composed of side chains that can assume only a limited number of conformations (*24*), efficient packing must be maintained without steric clashes. How important are hydrophobicity, volume, and steric complementarity in determining whether a given sequence can form an acceptable core? Each factor is essential in a physical sense, as a stable core is probably unable to tolerate unsatisfied hydrogen bonding groups, large holes, or steric overlaps (*25*). However, in an informational sense, these factors are not equivalent. For example, in experiments in which three core residues of λ repressor were mutated simultaneously, volume was a relatively unimportant informational constraint because three-quarters of all possible combinations of the 20 naturally occurring amino acids had volumes within the range tolerated in the core, and yet most of these sequences were unacceptable (*20*). In contrast, of the sequences that contained only

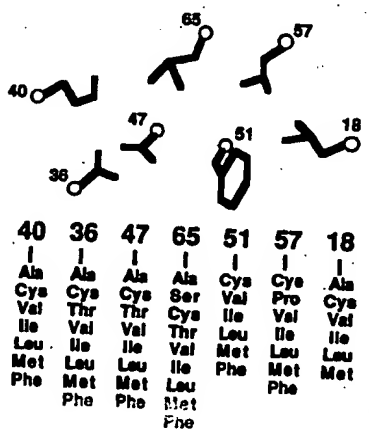

Fig. 2. Amino acid substitutions allowed in the core of λ repressor. The wild-type side chains are shown pictorially in the approximate orientation seen in the crystal structure (*43*). The lists of allowed substitutions at each position are shown below the wild-type side chains. These substitutions were identified by randomly mutating one to four residues at a time by using a cassette method and applying a functional selection (*20*). Not all substitutions are allowed in every sequence background.

the appropriate hydrophobic residues, a significant fraction were acceptable. Hence, the hydrophobicity of a sequence contains more information about its potential acceptability in the core than does the total side chain volume. Steric compatibility was intermediate between volume and hydrophobicity in informational importance.

The Informational Importance of Surface Sites

We have noted that many surface sites can tolerate a wide variety of side chains, including hydrophilic and hydrophobic residues. This result might be taken to indicate that surface positions contain little structural information. However, Bashford *et al.*, in an extensive analysis of globin sequences (*4*), found a strong bias against large hydrophobic residues at many surface positions. At one level, this may reflect constraints imposed by protein solubility, because large patches of hydrophobic surface residues would presumably lead to aggregation. At a more fundamental level, protein folding requires a partitioning between surface and buried positions. Consequently, to achieve a unique native state without significant competition from other conformations, it may be important that some sites have a decided preference for exterior rather than interior positions. As a result, many surface sites can accept hydrophobic residues individually, but the surface as a whole can probably tolerate only a moderate number of hydrophobic side chains.

Identification of Residue Roles from Sets of Sequences

Often, a protein of interest is a member of a family of related sequences. What can we infer from the pattern of allowed substitutions at positions in sets of aligned sequences generated by genetic or phylogenetic methods? Residue positions that can accept a number of different side chains, including charged and highly polar residues, are almost certain to be on the protein surface. Residue positions that remain hydrophobic, whether variable or not, are likely to be buried within the structure. In Fig. 3, those residue positions in λ repressor that can accept hydrophilic side chains are shown in orange and those that cannot accept hydrophilic side chains are shown in green. The obligate hydrophobic positions define the core of the structure, whereas positions that can accept hydrophilic side chains define the surface.

Functionally important residues should be conserved in sets of ıctive sequences, but it is not possible to decide whether a side chain s functionally or structurally important just because it is invariant or ɔnserved. To make this distinction requires an independent assay of ɔrotein folding. The ability of a mutant protein to maintain a stably ɔlded structure can often be measured by biophysical techniques, ɣ susceptibility to intracellular proteolysis (*26*), or by binding to ntibodies specific for the native structure (*27, 28*). In the latter ıses, it is possible to screen proteins in mutated clones for the ɔility to fold even if these proteins are inactive. Sets of sequences ıat allow formation of a stable structure can then be compared to ıe sets that allow both folding and function, with the active site or nding residues being those that are variable in the set of stable oteins but invariant in the set of functional proteins. The DNA-nding residues of Arc repressor were identified by this method (*8*). ıe receptor-binding residues of human growth hormone were also :ntified by comparing the stabilities and activities of a set of ıtant sequences (*28*). However, in this case, the mutants were ıerated as hybrid sequences between growth hormone and related :mones with different binding specificities.

Implications for Structure Prediction

At present, the only reliable method for predicting a low-resolution tertiary structure of a new protein is by identifying sequence similarity to a protein whose structure is already known (*29, 30*). However, it is often difficult to align sequences as the level of sequence similarity decreases, and it is sometimes impossible to detect statistically significant sequence similarity between distantly related proteins. Because the number of known sequences is far greater than the number of known structures, it would be advantageous to increase the reach of the available structural information by improving methods for detecting distant sequence relations and for subsequently aligning these sequences based on structural principles. In a normal homology search, the sequence database is scanned with a single test sequence, and every residue must be weighted equally. However, some residues are more important than others and should be weighted accordingly. Moreover, certain regions of the protein are more likely to contain gaps than others. Both kinds of information can be obtained from sequence sets, and several techniques have

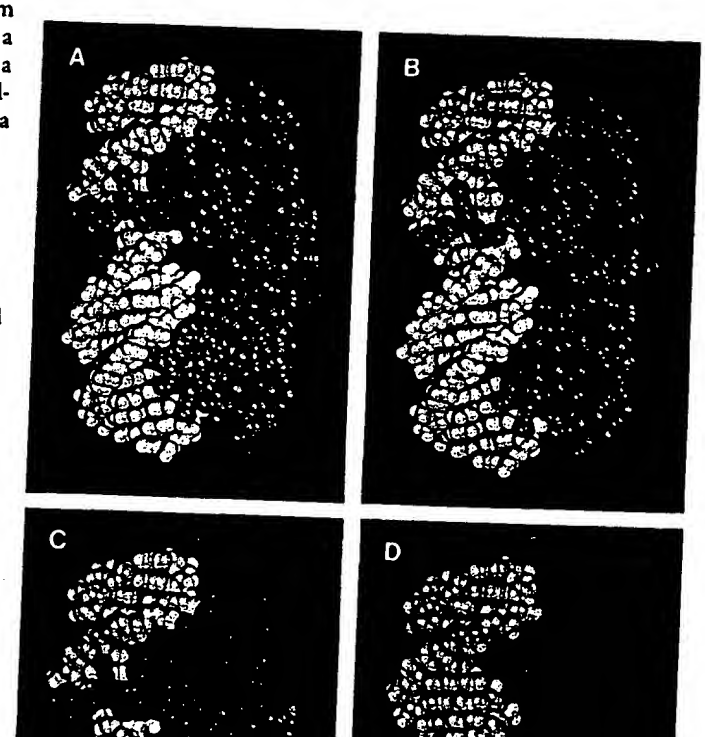

Fig. 3. Tolerance of positions in the NH_2-terminal domain of λ repressor to hydrophilic side chains. The complex (*43*) of the repressor dimer (blue) and operator DNA (white) is shown. In (**A**), positions that can tolerate hydrophilic side chains are shown in orange. The same side chains are shown in (**B**) without the remaining protein atoms. In (**C**), positions that require hydrophobic or neutral side chains are shown in green. These side chains are shown in (**D**) without the remaining protein atoms. About three-fourths of the 92 side chains in the NH_2-terminal domain are included in both (B) and (D). The remaining positions have not been tested. Data are from (*9, 14, 20, 27, 44*).

been used to combine such information into more appropriately weighted sequence searches and alignments (*31*). These methods were used to align the sequences of retroviral proteases with aspartic proteases, which in turn allowed construction of a three-dimensional model for the protease of human immunodeficiency virus type 1 (*29*). Comparison with the recently determined crystal structure of this protein revealed reasonable agreement in many areas of the predicted structure (*32*).

The structural information at most surface sites is highly degenerate. Except for functionally important residues, exterior positions seem to be important chiefly in maintaining a reasonably polar surface. The information contained in buried residues is also degenerate, the main requirement being that these residues remain hydrophobic. Thus, at its most basic level, the key structural message in an amino acid sequence may reside in its specific pattern of hydrophobic and hydrophilic residues. This is meant in an informational sense. Clearly, the precise structure and stability of a protein depends on a large number of detailed interactions. It is possible, however, that structural prediction at a more primitive level can be accomplished by concentrating on the most basic informational aspects of an amino acid sequence. For example, amphipathic patterns can be extracted from aligned sets of sequences and used, in some cases, to identify secondary structures.

If a region of secondary structure is packed against the hydrophobic core, a pattern of hydrophobic residues reflecting the periodicity of the secondary structure is expected (*33, 34*). These patterns can be obscured in individual sequences by hydrophobic residues on the protein surface. It is rare, however, for a surface position to remain hydrophobic over the course of evolution. Consequently, the amphipathic patterns expected for simple secondary structures can be much clearer in a set of related sequences (*6*). This principle is illustrated in Fig. 4, which shows helical hydrophobic moment plots for the Antennapedia homeodomain sequence (Fig. 4A) and for a composite sequence derived from a set of homologous homeodomain proteins (Fig. 4B) (*35*). The hydrophobic moment is a simple measure of the degree of amphipathic character of a sequence in a given secondary structure (*34*). The amphipathic character of the three α-helical regions in the Antennapedia protein (*36*) is clearly revealed only by the analysis of the combined set of homeodomain sequences. The secondary structure of Arc repressor, a small DNA-binding protein, was recently predicted by a similar method (*8*) and confirmed by nuclear magnetic resonance studies (*37*).

The specific pattern of hydrophobic and hydrophilic residues in an amino acid sequence must limit the number of different structures a given sequence can adopt and may indeed define its overall fold. If this is true, then the arrangement of hydrophobic and hydrophilic residues should be a characteristic feature of a particular fold. Sweet and Eisenberg have shown that the correlation of the pattern of hydrophobicity between two protein sequences is a good criterion for their structural relatedness (*38*). In addition, several studies indicate that patterns of obligatory hydrophobic positions identified from aligned sequences are distinctive features of sequences that adopt the same structure (*4, 29, 38, 39*). Thus, the order of hydrophobic and hydrophilic residues in a sequence may actually be sufficient information to determine the basic folding pattern of a protein sequence.

Although the pattern of sequence hydrophobicity may be a characteristic feature of a particular fold, it is not yet clear how such patterns could be used for prediction of structure de novo. It is important to understand how patterns in sequence space can be related to structures in conformation space. Lau and Dill have approached this problem by studying the properties of simple sequences composed only of H (hydrophobic) and P (polar) groups on two-dimensional lattices (*40*). An example of such a representation is shown in Fig. 5. Residues adjacent in the sequence must occupy adjacent squares on the lattice, and two residues cannot occupy the same space. Free energies of particular conformations are evaluated with a single term, an attraction of H groups. By considering chains of ten residues, an exhaustive conformational search for all 1024 possible sequences of H and P residues was possible. For longer sequences only a representative fraction of the allowed sequence or conformation space could be explored. The significant results were as follows: (i) not all sequences can fold into a "native" structure and only a few sequences form a unique native structure; (ii) the probability that a sequence will adopt a unique native structure increases with chain length; and (iii) the native states are compact, contain a hydrophobic core surrounded by polar residues, and contain significant secondary structure. Although the gap between these two-dimensional simulations and three-dimensional structures is large, the use of simple rules and sequence representations yields results similar to those expected for real proteins. Three-dimensional lattice methods are also beginning to be developed and evaluated (*41*).

Summary

There is more information in a set of related sequences than in a single sequence. A number of practical applications arise from an analysis of the tolerance of residue positions to change. First, such information permits the evaluation of a residue's importance to the function and stability of a protein. This ability to identify the essential elements of a protein sequence may improve our understanding of the determinants of protein folding and stability as well as protein function. Second, patterns of tolerance to amino acid substitutions of varying hydrophilicity can help to identify residues likely to be buried in a protein structure and those likely to occupy

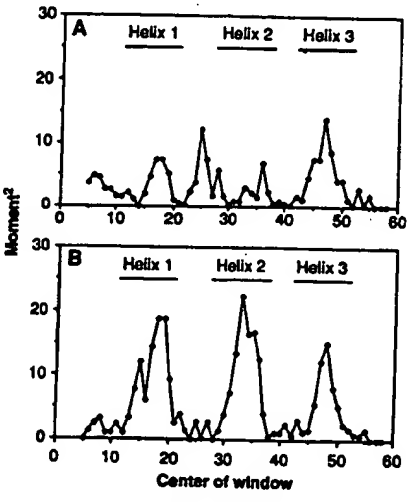

Fig. 4. Helical hydrophobic moments calculated by using (**A**) the Antennapedia homeodomain sequence or (**B**) a set of 39 aligned homeodomain sequences (*35*). The bars indicate the extent of the helical regions identified in nuclear magnetic resonance studies of the Antennapedia homeodomain (*36*). To determine hydrophobic moments, residues were assigned to one of three groups: H1 (high hydrophobicity = Trp, Ile, Phe, Leu, Met, Val, or Cys); H2 (medium hydrophobicity = Tyr, Pro, Ala, Thr, His, Gly, or Ser); and H3 (low hydrophobicity = Gln, Asn, Glu, Asp, Lys, or Arg). For the aligned homeodomain sequences, the residues at each position were sorted by their hydrophobicity by using the scale of Fauchere and Pliska (*45*). Arg and Lys were not counted unless no other residue was found at the position, because they contain long aliphatic side chains and can thereby substitute for nonpolar residues at some buried sites. To account for possible sequence errors and rare exceptions, the most hydrophilic residue allowed at each position was discarded unless it was observed twice. The second most hydrophilic residue was then chosen to represent the hydrophobicity of each position. An eight-residue window was used and the vectors projected radially every 100°. The vector magnitudes were assigned a value of 1, 0, or −1 for positions where the hydrophobicity group was H1, H2, or H3, respectively.

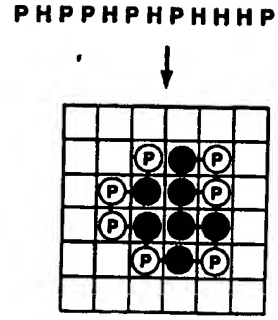

Fig. 5. A representation of one compact conformation for a particular sequence of H and P residues on a two-dimensional square lattice. [Adapted from (*40*), with permission of the American Chemical Society]

surface positions. The amphipathic patterns that emerge can be used to identify probable regions of secondary structure. Third, incorporating a knowledge of allowed substitutions can improve the ability to detect and align distantly related proteins because the essential residues can be given prominence in the alignment scoring.

As more sequences are determined, it becomes increasingly likely that a protein of interest is a member of a family of related sequences. If this is not the case, it is now possible to use genetic methods to generate lists of allowed amino acid substitutions. Consequently, at least in the short term, it may not be necessary to solve the folding problem for individual protein sequences. Instead, information from sequence sets could be used. Perhaps by simplifying sequence space through the identification of key residues, and by simplifying conformation space as in the lattice methods, it will be possible to develop algorithms to generate a limited number of trial structures. These trial structures could then, in turn, be evaluated by further experiments and more sophisticated energy calculations.

REFERENCES AND NOTES

1. C. J. Epstein, R. F. Goldberger, C. B. Anfinsen, *Cold Spring Harbor Symp. Quant. Biol.* **28**, 439 (1963); C. B. Anfinsen, *Science* **181**, 223 (1973).
2. R. E. Dickerson, *Sci. Am.* **242**, 136 (March 1980).
3. M. D. Hampsey, G. Das, F. Sherman, *FEBS Lett.* **231**, 275 (1988).
4. D. Bashford, C. Chothia, A. M. Lesk, *J. Mol. Biol.* **196**, 199 (1987).
5. A. M. Lesk and C. Chothia, *ibid.* **136**, 225 (1980).
6. M. F. Perutz, J. C. Kendrew, H. C. Watson, *ibid.* **13**, 669 (1965).
7. C. Chothia and A. M. Lesk, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 399 (1965).
8. J. U. Bowie and R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2152 (1989).
9. J. F. Reidhaar-Olson and R. T. Sauer, *Science* **241**, 53 (1988); *Proteins Struct. Funct. Genet.*, in press.
10. D. Shortle, *J. Biol. Chem.* **264**, 5315 (1989).
11. J. H. Miller *et al.*, *J. Mol. Biol.* **131**, 191 (1979).
12. S. Sprang *et al.*, *Science* **237**, 905 (1987); C. S. Craik, S. Roczniak, C. Largman, W. J. Rutter, *ibid.*, p. 909.
13. H. C. M. Nelson and R. T. Sauer, *J. Mol. Biol.* **192**, 27 (1986).
14. M. H. Hecht, J. M. Sturtevant, R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5685 (1984).
15. T. Alber, D. Sun, J. A. Nye, D. C. Muchmore, B. W. Matthews, *Biochemistry* **26**, 3754 (1987).
16. D. Shortle and A. K. Meeker, *Proteins Struct. Funct. Genet.* **1**, 81 (1986).
17. A. M. Lesk and C. Chothia, *J. Mol. Biol.* **160**, 325 (1982).
18. W. R. Taylor, *ibid.* **188**, 233 (1986).
19. W. Kauzmann, *Adv. Protein Chem.* **14**, 1 (1959); R. L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8069 (1986).
20. W. A. Lim and R. T. Sauer, *Nature* **339**, 31 (1989); in preparation.
21. Lesk and Chothia (*5*) have argued that a protein core composed solely of hydrogen-bonded residues would also be inviable on evolutionary grounds, as a mutational change in one core residue would require compensating changes in any interacting residue or residues to maintain a stable structure.
22. T. M. Gray and B. W. Matthews, *J. Mol. Biol.* **175**, 75 (1984); E. N. Baker and R. E. Hubbard, *Prog. Biophys. Mol. Biol.* **44**, 97 (1984).
23. F. M. Richards, *J. Mol. Biol.* **82**, 1 (1974).
24. J. W. Ponder and F. M. Richards, *ibid.* **193**, 775 (1987).
25. J. T. Kellis, Jr., K. Nyberg, A. R. Fersht, *Biochemistry* **28**, 4914 (1989); W. S. Sandberg and T. C. Terwilliger, *Science* **245**, 54 (1989).
26. A. A. Pakula and R. T. Sauer, *Proteins Struct. Funct. Genet.* **5**, 202 (1989).
27. B. C. Cunningham and J. A. Wells, *Science* **244**, 1081 (1989); R. M. Breyer and R. T. Sauer, *J. Biol. Chem.* **264**, 13348 (1989).
28. B. C. Cunningham, P. Jhurani, P. Ng, J. A. Wells, *Science* **243**, 1330 (1989).
29. L. H. Pearl and W. R. Taylor, *Nature* **329**, 351 (1987).
30. W. J. Brown *et al.*, *J. Mol. Biol.* **42**, 65 (1969); J. Greer, *ibid.* **153**, 1027 (1981); J. M. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 99 (1988).
31. W. R. Taylor, *Protein Eng.* **2**, 77 (1988).
32. M. A. Navia *et al.*, *Nature* **337**, 615 (1989).
33. M. Schiffer and A. B. Edmundson, *Biophys. J.* **7**, 121 (1967); V. I. Lim, *J. Mol. Biol.* **88**, 857 (1974); *ibid.*, p. 873.
34. D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Nature* **299**, 371 (1982); D. Eisenberg, D. Schwarz, M. Komaromy, R. Wall, *J. Mol. Biol.* **179**, 125 (1984); D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 140 (1984).
35. T. R. Burglin, *Cell* **53**, 339 (1988).
36. G. Otting *et al.*, *EMBO J.* **7**, 4305 (1988).
37. J. N. Breg, R. Boelens, A. V. E. George, R. Kaptein, *Biochemistry* **28**, 9826 (1989); M. G. Zagorski, J. U. Bowie, A. K. Vershon, R. T. Sauer, D. J. Patel, *ibid.*, p. 9813.
38. R. M. Sweet and D. Eisenberg, *J. Mol. Biol.* **171**, 479 (1983).
39. J. U. Bowie, N. D. Clarke, C. O. Pabo, R. T. Sauer, *Proteins Struct. Funct. Genet.*, in preparation.
40. K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
41. A. Sikorski and J. Skolnick, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2668 (1989); A. Kolinski, J. Skolnick, R. Yaris, *Biopolymers* **26**, 937 (1987); D. G. Covell and R. L. Jernigan, *Biochemistry*, in press.
42. B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971).
43. S. R. Jordan and C. O. Pabo, *Science* **242**, 893 (1988).
44. R. M. Breyer, thesis, Massachusetts Institute of Technology, Cambridge (1988).
45. J.-L. Fauchere and V. Pliska, *Eur. J. Med. Chem.-Chim. Ther.* **18**, 369 (1983).

46. We thank C. O. Pabo and S. Jordan for coordinates of the NH_2-terminal domain of λ repressor and its operator complex. We also thank P. Schimmel for the use of his graphics system and J. Burnbaum and C. Francklyn for assistance. Supported in part by NIH grant AI-15706 and predoctoral grants from NSF (J.R.-O.) and Howard Hughes Medical Institute (W.A.L.).